Lazzarini N, Runhaar J, Bay-Jensen AC, Thudium CS, Bierma-Zeinstra SMA, Henrotin Y, Bacardit J.

A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women.

***Osteoarthritis and Cartilage* 2017**

DOI: https://doi.org/10.1016/j.joca.2017.09.001

A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women

Lazzarini N[1,2], Runhaar J[2,3], Bay-Jensen AC[2,4], Thudium CS[2,4,] Bierma-Zeinstra SMA[2,3,5], Henrotin Y[2,6,7], Bacardit J[1,2*]

1) ICOS research group, School of Computing Science, Newcastle University, UK
2) D-BOARD Consortium, an FP7 programme by the European Committee
3) Erasmus University Medical Center Rotterdam, the Netherlands, dept. of General Practice
4) Nordic Bioscience, Copenhagen, Denmark
5) Erasmus University Medical Center Rotterdam, the Netherlands, dept. of Orthopedics
6) University of Liège, Belgium
7) Artialis SA, Liège, Belgium

* corresponding author
Email address: jaume.bacardit@newcastle.ac.uk
School of Computing
Urban Sciences Building,
Newcastle University, 1 Science Square, Science Central,
Newcastle upon Tyne,
NE4 5TG, UK.
United Kingdom

Abstract

Objective: Knee osteoarthritis (OA) is among the higher contributors to global disability. Despite its high prevalence, currently, there is no cure for this disease. Furthermore, the available diagnostic approaches have large precision errors and low sensitivity. Therefore, there is a need for new biomarkers to correctly identify early knee OA.

Study design and setting: We have created a machine learning based pipeline to identify small models (having few variables) that predict the 30-months incidence of knee OA (using multiple clinical and structural OA outcome measures) in overweight middle-aged women without knee OA at baseline. The data included clinical variables, food and pain questionnaires, biochemical markers and imaging-based information.
Results: All the models showed high performance (AUC > 0.7) while using only a few variables. We identified both the importance of each variable within the models as well its direction. Finally, we compared the performance of two models with the state-of-the-art approaches available in the literature.
Conclusions: We showed the potential of applying machine learning to generate predictive models for the knee OA incidence. Imaging-based information were found particularly important in the proposed models. Furthermore, our analysis confirmed the relevance of known biochemical markers for knee OA. Overall, we propose 5 highly predictive small models that can be possibly adopted for an early prediction of knee OA.

Keywords: knee osteoarthritis, machine learning, incidence, prediction

1 Introduction

Nowadays, knee OA is mainly diagnosed using clinical and radiographic changes generated by structural damages that occur late in the disease progression. In general, these techniques have a relatively large precision error and low sensitivity (1). Given the limitations of these imaging-based biomarkers (also known as “dry”), there is an increased need for identifying new and sensitive biochemical biomarkers (also called “wet”), other dry biomarkers (such as coming from MRI), or a combination of both that can detect early OA before structural damages and established clinical OA develop. Recently, several new approaches have been presented to tackle the lack of early knee OA detection (2). The levels of serum COMP have been correlated with the development of knee OA (3), the incidence of clinical knee OA among middle-aged overweight and obese women has been linked with the baseline fibulin-3 concentrations (4) and it was shown to be negatively associated with the baseline concentration of COLL2-1NO2 (5). Finally, adipokines are suggested as predictive biomarkers for early onset post-traumatic knee OA (6).

Machine learning has already been used to elucidate the underlying biological processes related to OA (7–9). In the present study, we use a pipeline (a set of machine learning-based computational procedures) to analyse a cohort of women at high-risk for knee OA development with the aim of identifying novel biomarkers that can contribute to the early detection of knee OA. The core of this pipeline is represented by RGIFE (10), a machine learning based heuristic designed to generate small yet highly predictive models from complex biomedical data. Different than many presented works, we used 5 distinct outcome measures of incident knee OA to generate small predictive models (at most 8 variables) that provide high classification performance. Furthermore, we investigate the contribution and the direction of each variable within the predictive models. Finally, we contrast the performance of the proposed models with state-of-the-art approaches available in the specialised literature. Our method results effective in identify relevant factors that are related with knee OA incidence. More importantly, although focused on the analysis of data associated with knee OA, the proposed methodology is generic enough to be applied to a wide variety of biomedical data. It possesses the potential to discover and analyse the role of novel biomarkers for many different conditions and can help in understand their mechanisms.

2 Method

2.1 Dataset and individuals

The data used in this work came from the PROOF study; a preventive randomised controlled trial including 407 middle-aged women with a BMI ≥27 kg/m2 free of clinical knee OA at baseline (11). After 30 months, the preventive effects of a diet & exercise program and of oral glucosamine sulphate were evaluated. Since no intervention effects were found, data were here treated as a cohort, with 5 different outcome measures of incident knee OA after 30 months:

- incidence of ‘combined radiographic and clinical ACR-criteria’
- incidence of frequent knee pain
- lateral JSN of ≥ 1.0 mm
- medial JSN of ≥ 1.0 mm
- incidence of KL ≥ 2

Each individual was characterised by the value of 186 heterogeneous baseline variables (the full list of variables is available in the Supplementary Material). Data were derived from baseline questionnaires (including demographics, menopausal status, knee complaints, physical activity level, quality of life, habitual nutritional intake, and KOOS questionnaire), radiographs (for obtaining baseline KL grade, medial alignment angle, and knee joint shape using active shape modelling), MR images (scored with semi-quantitative MOAKS system (12) and used to define MRI OA (13)), physical examination (including pain upon palpation of knee structures, crepitus, presence of Heberden's nodes, blood pressure, knee laxity and range of motion, warmth of the knee joint, waist circumference, and skinfolds for fat percentage calculation), and biochemical markers from serum and urine (fibulin3-1, fibulin3-2 and fibulin3-3 (4), COLL2-1NO2 (5), and C1M and C2M (14). A detailed description of the acquisition of non-biochemical variables is given elsewhere (11). Separate analysis was performed for the different definition of knee OA. From now onward we will refer to the 5 definitions, and the associated analysis, as: ACR criteria, Knee pain, Lateral JSN, Medial JSN and KL incidence.

2.2 A machine learning pipeline for the generation of small predictive models

2.2.1 The RGIFE heuristic

The aim of this study was to identify biomarkers that can discriminate between *incidence* and *non-incidence* for 5 different outcomes measures of knee OA. *RGIFE* (Ranked Guided Iterative Feature Elimination) is a machine learning heuristic able to select few biomarkers with high predictive power (10). RGIFE is an iterative based heuristic, it discards features if their removal does not decrease the overall predictive performance of the computational model. By iteratively repeating the reduction process and performing other optimisation decisions, the method has been proven to select small sets of variables with high predictive power when analysing complex transcriptomics dataset (10). In fact, RGIFE has been designed to deal with challenging and difficult biomedical data, often characterised by a small number of samples that are likely to be defined in a high-dimensional space (that is described by thousands of variables). With specific optimisation techniques (e.g. oversampling, cost-sensitive learning, etc.), RGIFE can converge to reduced panels of biomarkers.

In our analysis, RGIFE used the random forest algorithm (15) to build the predictive models. A random forest consists of a collection of simple decision trees, where each of them is generated using a random (different) subsets of training samples and variables. The prediction for each individual is made based on the majority vote of the set of trees. That is, each sample is assigned to the class predicted by the majority of the decision trees. The PROOF data represented a difficult task from a machine learning point of view due to: 1) presence of many missing values and 2) imbalance distribution of the samples (much more *non-incidence* than *incidence*). The missing values were imputed using the K-means algorithm, while the imbalanced class distribution problem was tackled using the SPIDER oversampling method (16). For both methods we used the implementation in the KEEL machine learning package (17).

2.2.2 Identification of the best model

RGIFE is a flexible and fine tuneable algorithm, we used 30 different configurations to perform a full search in the space of all the optimal solutions (set of biomarkers) for each OA definition. The configurations differed in terms of maximum depth the random forest and misclassification costs (penalisation when misclassifying *incidence* samples during the learning phase).

The best performing configuration was selected with a 10-fold cross-validation, a typical approach used in machine learning to evaluate the performance of a predictive. A *n*-fold cross-validation (in our analysis n=10) randomly divides the dataset into *n* equally-sized disjoint subsets (folds), each of them having the same distribution of positive and negative samples as in the complete dataset. In turn, each set is used as test set while the remaining n-1 are used as training set. By calculating the performance obtained using the test sets we can assess how the model will generalise to an independent dataset. In our analysis the 10-fold cross-validation was repeated 10 times to minimise the bias introduced by the data being split into training and test set. Afterwards, using RGIFE on the complete dataset, we identified the best performing models (highest AUC calculated with a new 10 x 10-fold cross-validation) with at most 10 variables. The overall analytical process is presented in Figure 1. Because of this pipeline, in the results section, we will report two different AUC values, namely AUC-CV and AUC-Full. AUC-CV refers to the AUC obtained by the best performing configuration using a 10 x 10-fold cross-validation. AUC-Full, indicates the

performance, calculated using a new 10 x 10-fold cross-validation, of the best selected model generated when considering to the complete set of samples (with RGIFE using the best performing configuration).

Finally, we used a permutation test to assess how much better than random our models are. We contrasted the performance of 100 random models, generated from 100 permutated datasets (where the labels *incidence* and *non-incidence*) were randomly assigned) to the AUC obtained from the original data. An empirical p-value was estimated by counting the number of times in which a model provides better performance when trained with random data rather than the original data.

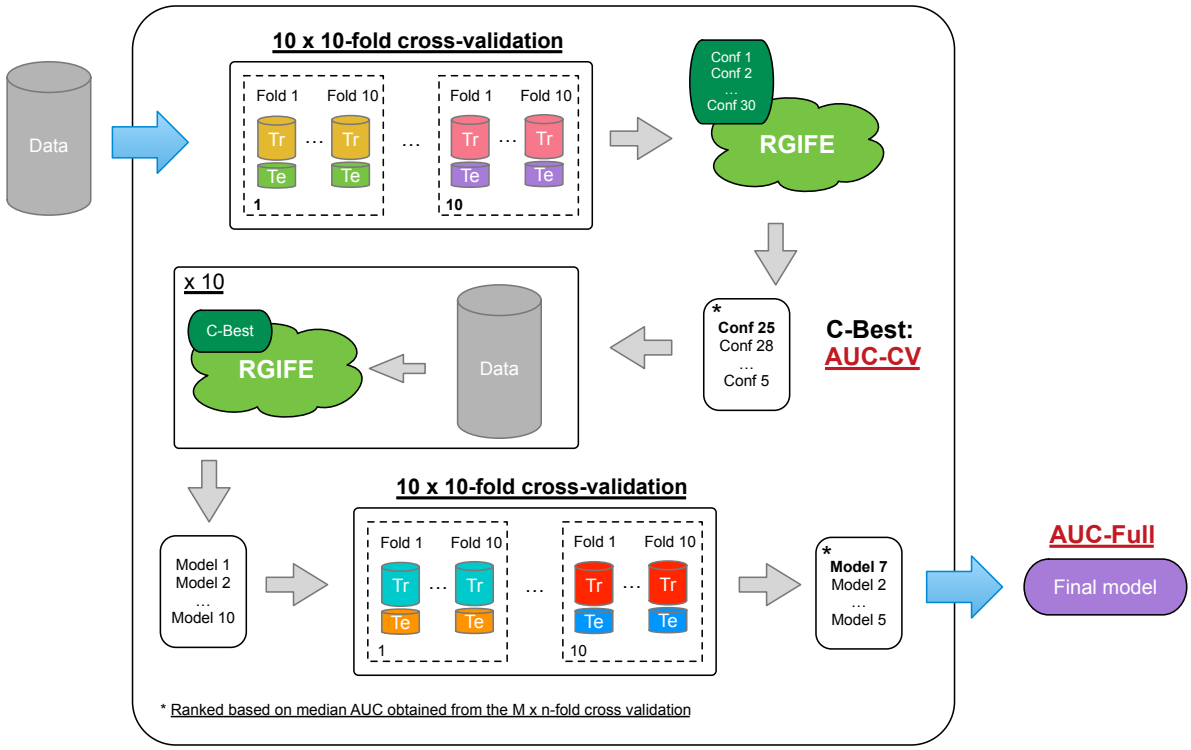

Fig. 1 Analytic pipeline employed for the identification of the best predictive models. First, using a 10 x 10-fold cross-validation, the best performing configuration of RGIFE is identified. The predictive performance of the best configuration is indicated as AUC-CV. Then, using the selected set of parameters, RGIFE is applied 10 times (due to its stochastic behaviour) to the whole set of samples. Finally, the 10 generated models are ranked using the median AUC obtained with a new 10 x 10-fold cross-validation. The top ranked model, whose performance is indicated as AUC-Full, is selected for the biomarker identification

2.2.3 Model interpretation

We analysed the role of the variables, within the models, by assessing both

their (additive) value and their association with the outcome (incidence of knee OA in this instance). The additive value was obtained from the generation of sub-models of decremental sizes, named *decremental analysis*. That is, from the original set of variables, we iteratively identified the one whose removal caused the smallest drop of AUC, hence contributing the less within the prediction task. The variable direction indicates how the value assumed by a variable, influences the presence of the disease (condition). This was accomplished by performing a partial dependence analysis, a method to visualise the partial relationship between the outcome and the predictive variables (18). The method evaluates the variations in predictions when the values of the variables are changed, thus it determines the relationship of the biomarkers with the outcome measure (see Section 2 of the Supplementary Material for a more detailed description).

3 Results

Out of all the subjects included in the PROOF study, 365 had follow-up data and were selected for the present study. The baseline characteristics of the individuals are presented in Table 1. A different total number of subjects was available for different knee OA outcome measures. The ACR criteria and chronic knee pain after 30 months occurred in 39 out 354 (11%) and in 51 out of 351 (15%) women respectively. The incidence of lateral JSN ≥ 1.0 mm was assessed in 41 (12%) out of 352 women, while medial JSN was seen in 38 (11%) women out of 352. Finally, the incidence of K&L ≥ 2 was measured in 27 (8%) out of 321 individuals. A supplementary analysis was performed by using an outcome measure that combines 4 of the outcome measures: ACR criteria, K&L score, lateral and medial JSN. The aim was to check if a combined measure could help in generating better models. The results of this analysis are available in Section 3 of the Supplementary Material.

Table 1 Baseline characteristics of the included subjects (N = 365)

	Mean ± SD or percentage
Age (yr)	55.7 ± 3.2
BMI (kg/m^2)	32.3 ± 4.3
Menopausal status	69%
Western etnicity	96%
Mild symptoms	45%
Physical activity (SQUASH score)	6900 ± 3700

K&L = 1	60%
K&L = 2	10%

3.1 Best predictive models

In Table 2, we report the models generated from the different OA definitions. The variables are grouped based on their source of information: OA measures (OA), clinical information (CI), Imaging-based data (IM), biochemical markers (BM), pain (PQ) and food questionnaire (FQ), all coming from the baseline assessments. Figure 2 shows the ROC curves generated from each of the generated models.

The best performing model was generated using the *KL incidence* OA outcome measure; it obtained an AUC-Full of 0.823 by using only 5 variables (smallest model). The second best performing model is associated with the *ACR criteria* and provided an AUC-Full of 0.788, however, it also contains the largest number of variables (8 in total). Finally, the *JSN* outcome measures led to the two lowest performance, respectively 0.731 for the lateral and 0.737 for the medial compartments. It is important to highlight how the rank of the models, created with different outcome measures, does not change while considering AUC-Full or AUC-CV.

Table 2 shows that all the models incorporate variables of diverse categories, each one is defined by at least four different categories of variables. The imaging-based variables are important for the *ACR criteria* model as well for *JSN lateral.* In particular, the Active Shape Models, statistical shape models constructed after placing 75 landmark points along the contours of the tibia, fibula, femur and medial femoral condyle on each radiograph and using Principal Component Analysis to create independent modes of shape variation. More details are provided in (19,20). OA measures play a relevant role in the prediction of knee OA development when considering the chronic pain as outcome measure. Food information appears in almost every model (*ACR criteria* is the exception), most variables are related to the fruit intake per week, while the number of biscuits (sugar) consumed per week is used by the *Knee pain* model. Finally, we notice the presence of biochemical markers already associated, from the literature, with the incidence of knee OA such as the concentration of C1M, C2M (14) and COLL2-1NO2 (5).

Table 2 Summary of the best models for each knee OA outcome measure. Variables are baseline measures divided according to the type of information provided: OA information (OA), clinical information (CI), Imaging-based information (IM), pain questionnaire (PQ) and food questionnaire (FQ). The AUC column includes the AUC-CV, the AUC-Full and the 95% Confidence Interval (Conf.Int.) of the AUC-Full.

OA measure	Biomarkers	Cat.	AUC-Full/ AUC-CV
ACR criteria	● KL grade ≥ 1 in one or both knees ● Maximal isometric quadriceps strength ● Mode 10, Mode 15, Mode 11 (Active Shape Modelling) ● Presence of knee pain in the last month, Difficulties when kneeling ● C2M concentration	OA CI IM PQ BM	0.788 / 0.692 Conf.Int. 0.712-0.863
Knee pain	● KL grade ≥ 1 in one or both knees, KL grade ≥ 2 in one or both knees, WOMAC function score ● Maximal isometric quadriceps strength ● Mode 11 (Active Shape Modelling) ● Difficulties when jumping ● Frequency biscuits (raisins) /week	OA CI IM PQ FQ	0.755 / 0.637 Conf.Int. 0.680-0.830
Lateral JSN	● Fat percentage ● Mode 1, Mode 10, Mode 11 (Active Shape Modelling) ● Frequency of fruits / week ● Concentration of Coll2-1NO2 adj. for creatinine	CI IM FQ BM	0.731 / 0.549 Conf.Int. 0.654-0.808
Medial JSN	● Quality of life, Nr. years since menopause, Waist circumference ● Mode 15 (Active Shape Modelling) ● Freq. bananas / week ● C1M concentration	CI IM FQ BM	0.737 / 0.539 Conf.Int. 0.659-0.814
KL incidence	● BMI, HbA1c concentration ● Presence of OA on MRI ● Grinding / clicking sound when moving the knee ● Frequency of apples and pears / week	CI IM PQ FQ	0.823 / 0.699 Conf.Int. 0.753-0.893

All the performance were found statistically significant with a p-value < 0.0001. None of the models, when trained with randomised data, obtained higher AUC than what can be obtained from the original versions of the data (reported in Table 2).

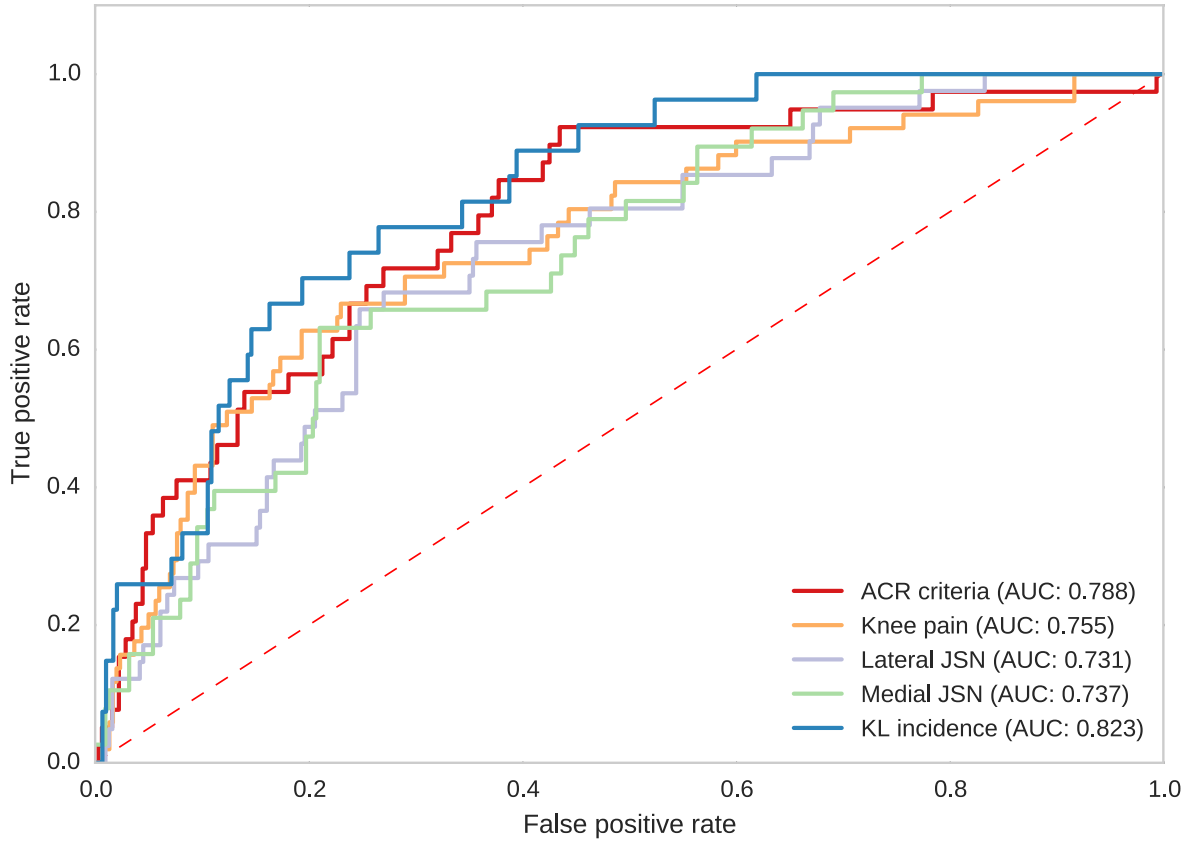

Fig. 2 ROC curves of the presented modes. The ROC curves generated by the best performing models using five knee OA outcome measures. The AUC values refer to the AUC-Full

3.2 Model interpretation

The models generated from different knee OA outcome measures were interpreted in terms of both additive value and direction of their components (variables).

The additive value of each variable, within the predictive models, was assessed performing a decremental analysis. In Figure 3 we show the results of the analysis when using the *Knee pain* and *KL incidence* models (the only two models comparable with the literature findings) The full set of results, model by model, is available in Section 1 of the Supplementary Material. The y-axis of Figure 3 represents the AUC values of each submodel, while the variables defining the submodels are represented over each bar. In white the least important variables are highlighted, that is the variables that are not present in the following submodels. The KL scores (whether at baseline is ≥ 1 or ≥ 2) are the most valuable information in the *Knee pain* model, by themselves, the two variables get a good AUC of 0.663. From the *KL incidence* measure models we see that the two largest decreases in performance are associated

with the removal of the fruit intake per week and the knee grinding/clicking information. As for the *Knee pain*, the one variable model (presence of OA on MRI at baseline) can lead to an interesting AUC of 0.617.

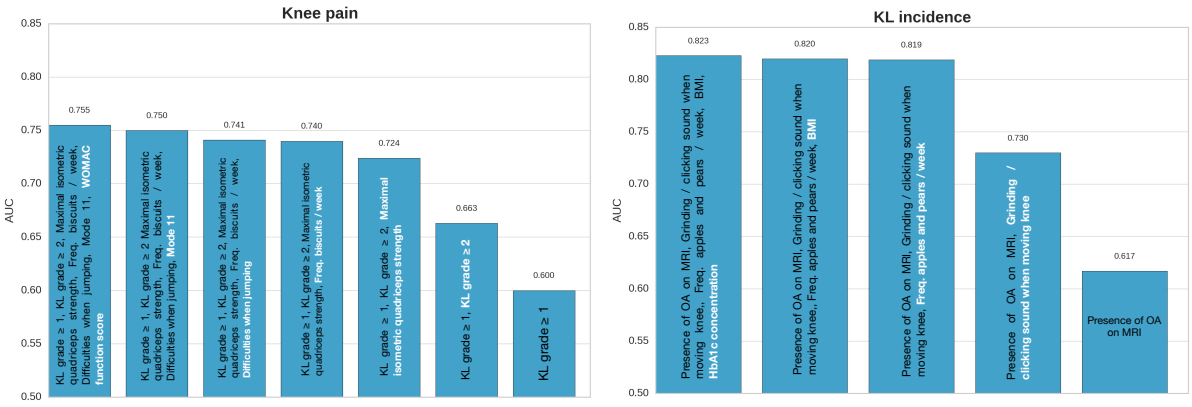

Fig. 3 Decremental models analysis. The decremental models generated using the *Knee pain* (left plot) and the *KL incidence* (right plot) measure for the knee OA definition. The variable highlighted in white represents the least, not present in the subsequent decremental model.

With the variable direction analysis, we determine if the change in intensity (value) of a variable corresponds to an increase of the probability to be affected by knee OA. That is, if a biomarker has a positive or a negative association with the OA outcome measure. Figure 4 shows the direction of two variables for *the Knee pain* and *KL incidence* models. Each data point corresponds to the average probability to belong to the positive (*incidence*) class. For binary variables (e.g. KL-grade ≥ 2) we report the distribution of the probabilities of the two possible values. As expected, a baseline KL grade ≥ 2 increases the probability for the knee OA incidence. We also see that difficulties when jumping (high pain) might indicate higher chance develop the condition. Conversely, at the bottom plots of Figure 4, is illustrated a negative association between the frequency of apples and pears eaten per week and the knee OA outcome measure. Finally, we observe how the analysis of MRI (at baseline) might provide insights about the development of knee OA in the individuals (overweight women in this instance) given its positive association.

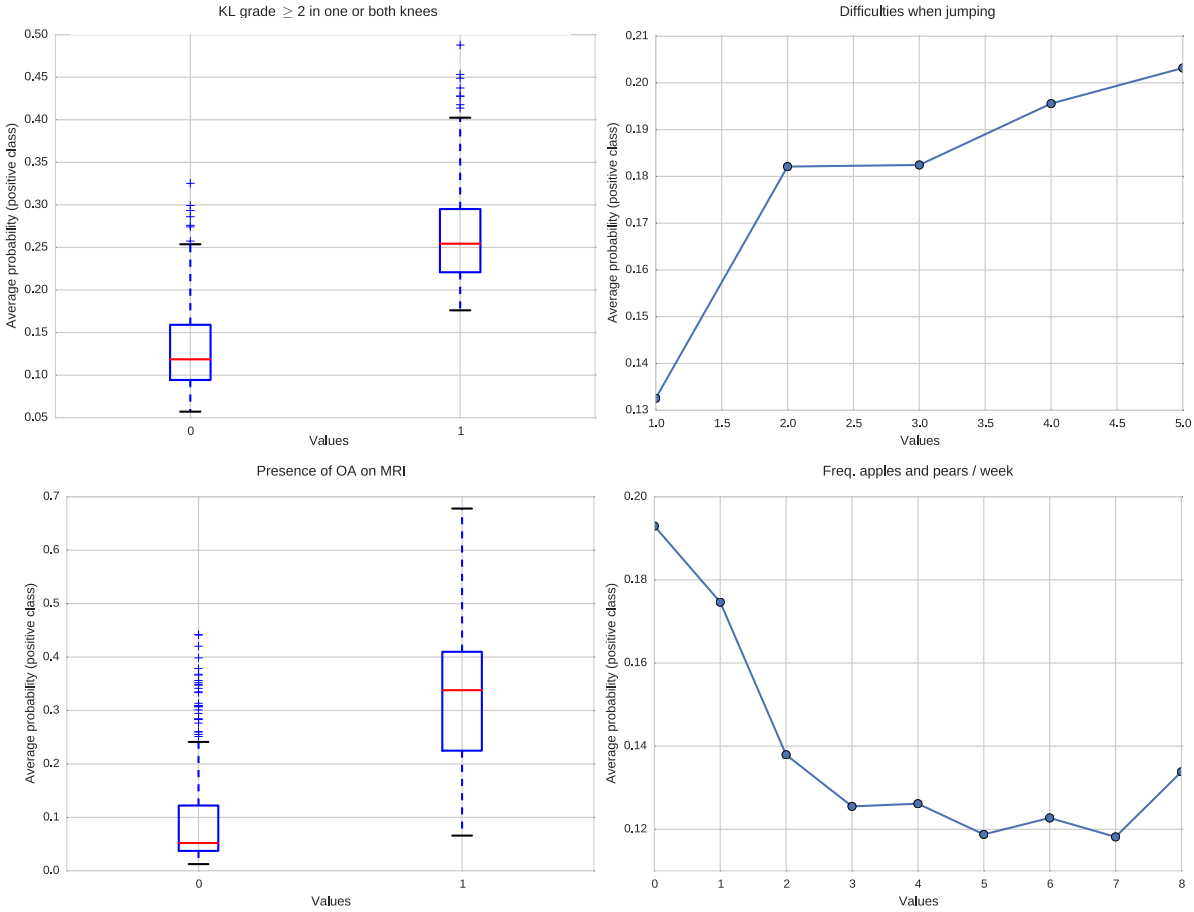

Fig. 4 Variable direction analysis. The top plots show the direction of the "*KL grade ≥ 2*" and "*Difficulties when jumping*" variables from the *Knee pain* model. The bottom plots illustrate the associations between the incidence of knee OA and the "*Presence of OA on MRI"* and "*Frequency of apples and pears eaten per week*" variables from the *KL incidence* model

4 Discussions

In this work, we developed 5 models that can be used to predict the incidence of knee OA in overweight and obese women. Differently than the traditional approaches, mostly based on the combination of univariate (where one variable at the time is analysed) and multivariate (where multiple variables are analysed together) logistic regression, we employed multivariate machine learning techniques. We showed that using a small subset of the available information (each model is defined by at most 8 variables) is possible to accurately predict the incidence of knee OA.

We searched the specialised literature to identify models that can be compared with the predictive models generated using our machine learning approach. In Table 3 and 4 we report a summary of the literature findings. We searched for models generated using data where the OA definition was similar to ours. Comparable models were

found only for the *Knee pain* and *KL incidence* outcome measures. For a fair comparison, we report, for each literature study, only the model AUC obtained using an internal validation (some studies also used external data to assess the performance). If multiple models (defined by different subset of variables) were available, we considered only the best performing. We want to highlight that these studies used a population that is different to the one associated with the PROOF study. Hence, the comparison of AUC values is just approximate. Nevertheless, such a comparison provides an idea if the presented models perform in line with the state-of-the-art literature. All the models found in the literature were generated with the same statistical approach: a univariate analysis followed by a multivariate logistic regression method. In (21–23) the validation protocol, employed calculate the AUC values, is not described. Therefore, we assume, as common practice for clinical studies, that the published AUCs are equivalent to our AUC-Full values. In (24) is mentioned that a "10-fold cross-validation strategy has been used as a feature selection strategy", however, is not specified if the reported AUCs are associated with such strategy. Our internal validation indicated an AUC of 0.823 for the *KL incidence* model and an AUC of 0.755 for the *Knee pain*.

Table 3 Summary of *the KL incidence* models identified in the specialised literature. JSW: Joint Space Width, B/L: baseline, F.U.: follow-up

Reference	AUC	OA definition	Attributes
(21)	0.790	KL grade < 2 B/L and KL ≥ 2 at F.U.	Gender, age, BMI, knee pain and KL score at B/L
(22)	0.690	KL grade < 2 B/L and KL ≥ 2 at F.U.	Gender, age, BMI, occupational risks, family osteoarthritis, previous knee injury
(23)	0.740	KL grade < 2 B/L and KL ≥ 2 at F.U. (5 year)	Gender, age, BMI, minimum JSW, osteophyte

Table 4 Summary of *the Knee pain* models identified in the specialised literature. JSW: Joint Space Width, B/L: baseline, F.U.: follow-up

Reference	AUC	OA definition	Attributes

(23)	0.600	Painful knee at B/L; painful knee at F.U (4 and 5 year)	Age, pain intensity, minimum JSW, osteophyte
(24)	0.623	Chronic right knee pain	Osteophytes (OARSI grades 0–3) femur medial compartment, Chondrocalcinosis (grades 0-1) medial compartment (data taken 1 year before the pain development)
(24)	0.620	Chronic right knee pain	Osteophytes (OARSI grades 0–3) femur medial compartment, Chondrocalcinosis (grades 0-1) medial compartment, Osteophytes (OARSI grades 0–3) femur lateral compartment (data taken 2 years before the pain development)

Table 3 shows that our model for the prediction of *KL incidence* obtains higher performance than the models available in the literature while using the same or fewer variables. A superior performance also emerged when comparing the *Knee pain* model (see Table 4). However, differently than for the *KL incidence*, our *Knee pain* model is always larger (7 attributes) and more heterogeneous (imaging-based information, food and pain questionnaire data, OA and clinical information).

Within our modes, imaging-based variables contains valuable information for the risk for knee OA incidence. Every model of Table 2 contains a variable from the imaging category (IM). In addition, such an importance is confirmed by Figure 4, where is illustrated a positive association between the presence of OA on MRI at baseline and the incidence of knee OA (using KL grade). Overall, these results suggest a need in re-evaluating a proper use of imaging information in primary care settings, especially when treating subjects at risk for future knee OA development. However, it needs to be stressed that not all imaging-based variables included in the final models are easy to obtain in primary care; such as the outcomes of statistical shape modelling (‘Mode x’ variables). This warrants further attention. Furthermore, we hope that the relevant role of MRI features for the prediction of knee OA incidence in our model might direct the design of new studies that focus on early detection or early treatment of knee OA among a high-risk group of overweight and obese women.

The models generated with different knee OA outcome measures do not share many biomarkers. This is primarily due to the lack of overlap between “incident” individuals across different OA definitions. The Venn diagram included in Section 5 of the Supplementary Material illustrates this. There is no common agreement between all the measures for any single incident individual. The largest similarity occurs between *ACR criteria* and *knee pain*, the two measures that also generated the most analogous predictive models (this occurs because the ACR criteria can only be positive in presence of knee pain). However, this a low overlap might also be a signed that potentially different factors play a role in different aspects of the disease (radiographic onset, clinical onset, etc.), hence each panel of biomarkers exists of unique factors. A more tailored analysis in future work will be necessary to thoroughly characterise the potential link between models.

The proposed biomarkers can be grouped into two predictive categories: early signs (e.g. pain while jumping) as well as risk factors (e.g. BMI or waist circumference). The analysis performed for this manuscript used all the data collected by the PROOF study without discarding or filtering any of the available variables, this was to avoid any loss of relevant information. However, in future, we would like to apply the same approach by considering either only the early signs or the risk factors. Such an approach will give us the opportunity to propose different models that can be applied according to the information available for each possible individual been tested for the prediction of knee OA.

Several variables based on “wet” biomarkers of extracellular matrix tissue turnover also provided information on the risk of knee OA incidence. Both ACR criteria, medial and lateral JSN based prediction models all contained blood or urine-based biomarkers. C1M and C2M seemed to be negatively associated with the incidence OA defined by ACR criteria and medial JSN, in a similar way, COLL2-1NO2 showed a negative relationship with incidence OA defined by lateral JSN. Previous studies have found correlations between OA severity and C1M (14) and a difference in C2M levels between OA and healthy subjects (25). Furthermore, the negative association between COLL2-1NO2 and OA incidence is in line with a previous study also performed using the PROOF cohort (5) with a different analytical approach (binary logistic regression), thus providing some degree of confirmation of both the relevance of the biochemical marker and the goodness of our methodology. These findings suggest that assessment of structural degradation products from the extracellular

matrix in body fluids may provide valuable information on the development OA and prediction of disease incidence in high-risk groups.

The presented models perform quite well in the discrimination of *incidence* and *non-incidence* knee OA among overweight and obese women. Their performance can be considered as “fair” (AUCs is between 0.70 - 0.80) and “good” (AUCs between 0.80-0.90). However, it will be necessary to evaluate the predictive power of each model using an independent set of individuals (external validation). Given that it will be extremely unlikely to obtain new validation data that will include all the variables employed by our models, the role of the decremental analysis will be fundamental. Based on the variables available in the new data, the performance of the best fitting submodel will be easily extracted from the decremental analysis, illustrated as in Figure 3, and compared with the AUCs obtained from the new external samples.

Acknowledgement: This work made use of the facilities of N8 HPC Centre of Excellence, provided and funded by the N8 consortium and EPSRC [EP/K000225/1]. The Centre is co-ordinated by the Universities of Leeds and Manchester.
We acknowledge the HPC facility at the School of Computing Science of Newcastle University for providing the necessary framework for the experiments.

† http://www.d-board.eu/dboard/index.aspx

Contributions: Conception and design of the study (NL, JB, JR). Performed the experiments (NL, JB). Analysis and interpretation of the data (NL, JB, JR, AB, CS, SZ, YH). Drafting of the article and final approved of the submitted version (NL, JB, JR, AB, CS, SZ, YH).

Conflict of interest: Yves Henrotin is CEO, President and Founder of Artialis SA (Spin-off company of the University of Liège, Belgium). Anne-Christine Bay-Jensen owns stocks in Nordic Bioscience.

Role of the funding source: This work was supported by the European Commission through the D-BOARD† Consortium funded by European Commission Framework 7 programme (EU FP7; HEALTH.2012.2.4.5-2, project number 305815,

Novel Diagnostics and Biomarkers for Early Identification of Chronic Inflammatory Joint Diseases).
This work was partly supported by a program grant of the Dutch Arthritis Foundation for their centre of excellence “Osteoarthritis in primary care”.
None of the study sponsor was involved with the design of the study, analysis and interpretation of data nor in the writing of the manuscript.

References

1. Wright RW, Boyce RH, Michener T, Shyr Y, McCarty EC, Spindler KP. Radiographs are not useful in detecting arthroscopically confirmed mild chondral damage. Clin Orthop Relat Res [Internet]. 2006;(442):245–51. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-33645079681&partnerID=40&md5=ef956150bde0e5c31364b945ccb73daf
2. Mobasheri A, Henrotin Y. Biomarkers of (osteo)arthritis. Biomarkers [Internet]. 2015;20(8):513–8. Available from: http://www.tandfonline.com/doi/full/10.3109/1354750X.2016.1140930
3. Kluzek S, Bay-Jensen A-C, Judge A, Karsdal MA, Shorthose M, Spector T, et al. Serum cartilage oligomeric matrix protein and development of radiographic and painful knee osteoarthritis. A community-based cohort of middle-aged women. Biomarkers [Internet]. 2015;20(8):557–64. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4819573&tool=pmcentrez&rendertype=abstract
4. Runhaar J, Sanchez C, Taralla S, Henrotin Y, Bierma-Zeinstra SMA. Fibulin-3 fragments are prognostic biomarkers of osteoarthritis incidence in overweight and obese women. Osteoarthr Cartil. 2016;24(4):672–8.
5. M.L. L, J. R, Y.E. H, M. VM, E.H. O, D. V, et al. COLL2-1NO2: A biomarker for early knee osteoarthritis? Osteoarthr Cartil [Internet]. 2014;22(2014):S77. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=71463888
6. Kluzek S, Arden NK, Newton J. Adipokines as potential prognostic biomarkers in patients with acute knee injury. Biomarkers [Internet]. 2015;20(8):519–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26006054%5Cnhttp://www.pubmedcentra

l.nih.gov/articlerender.fcgi?artid=PMC4819580

7. Ahmed U, Anwar A, Savage RS, Costa ML, Mackay N, Filer A, et al. Biomarkers of early stage osteoarthritis, rheumatoid arthritis and musculoskeletal health. Sci Rep [Internet]. 2015;5:9259. Available from: http://www.nature.com/articles/srep09259%5Cnhttp://www.nature.com/doifinder/10.1038/srep09259
8. Ashinsky BG, Coletta CE, Bouhrara M, Lukas VA, Boyle JM, Reiter DA, et al. Machine learning classification of OARSI-scored human articular cartilage using magnetic resonance imaging. Osteoarthritis Cartilage [Internet]. 2015;23(10):1704–12. Available from: http://www.sciencedirect.com/science/article/pii/S1063458415011851
9. Heard BJ, Rosvold JM, Fritzler MJ, El-Gabalawy H, Wiley JP, Krawetz RJ. A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers. J R Soc Interface [Internet]. 2014;11(97):20140428. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24920114%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4208376
10. Lazzarini N, Bacardit J. RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers. BMC Bioinformatics [Internet]. 2017;18(1):322. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28666416%5Cnhttp://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1729-2
11. Runhaar J, Van Middelkoop M, Reijman M, Willemsen S, Oei EH, Vroegindeweij D, et al. Prevention of Knee Osteoarthritis in Overweight Females: The First Preventive Randomized Controlled Trial in Osteoarthritis. Am J Med. 2015;128(8):888–895.e4.
12. Hunter DJ, Guermazi A, Lo GH, Grainger AJ, Conaghan PG, Boudreau RM, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). Osteoarthr Cartil. 2011;19(8):990–1002.
13. Hunter DJ, Arden N, Conaghan PG, Eckstein F, Gold G, Grainger A, et al. Definition of osteoarthritis on MRI: Results of a Delphi exercise. Osteoarthr Cartil. 2011;19(8):963–9.
14. Siebuhr AS, Petersen KK, Arendt-Nielsen L, Egsgaard LL, Eskehave T, Christiansen C, et al. Identification and characterisation of osteoarthritis

patients with inflammation derived tissue turnover. Osteoarthr Cartil. 2014;22(1):44–50.
15. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
16. Napierała K, Stefanowski J, Wilk S. Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In: Szczuka M, Kryszkiewicz M, Ramanna S, Jensen R, Hu Q, editors. Rough Sets and Current Trends in Computing: 7th International Conference, RSCTC 2010, Warsaw, Poland, June 28-30,2010 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 158–67.
17. Alcalá-Fdez J, Sánchez L, García S, Jesus MJ del, Ventura S, Garrell JM, et al. KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput. 2009;13(3):307–18.
18. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Elements. 2009;1:337–87.
19. Eggerding V, van Kuijk KSR, van Meer BL, Bierma-Zeinstra SM a, van Arkel ER a, Reijman M, et al. Knee shape might predict clinical outcome after an anterior cruciate ligament rupture. Bone Jt J [Internet]. 2014;96–B(6):737–42. Available from: http://www.bjj.boneandjoint.org.uk/content/96-B/6/737.abstract
20. Haverkamp DJ, Schiphof D, Bierma-Zeinstra SM, Weinans H, Waarsing JH. Variation in joint shape of osteoarthritic knees. Arthritis Rheum. 2011;63(11):3401–7.
21. Kerkhof HJ, Bierma-Zeinstra SM, Arden NK, Metrustry S, Castano-Betancourt M, Hart DJ, et al. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. Ann Rheum Dis. 2014;73(12):2116–21.
22. Zhang W, McWilliams DF, Ingham SL, Doherty S a, Muthuri S, Muir KR, et al. Nottingham knee osteoarthritis risk prediction models. Ann Rheum Dis. 2011;70(9):1599–604.
23. Kinds MB, Marijnissen ACA, Vincken KL, Viergever MA, Drossaers-Bakker KW, Bijlsma JWJ, et al. Evaluation of separate quantitative radiographic features adds to the prediction of incident radiographic osteoarthritis in individuals with recent onset of knee pain: 5-year follow-up in the {CHECK} cohort. Osteoarthr Cartil. 2012;20(6):548–56.
24. Galván-Tejada JI, Celaya-Padilla JM, Treviño V, Tamez-Peña JG. Multivariate

Radiological-Based Models for the Prediction of Future Knee Pain: Data from the {OAI}. Comput Math Methods Med. 2015;2015:1–10.

25. Bay-Jensen AC, Liu Q, Byrjalsen I, Li Y, Wang J, Pedersen C, et al. Enzyme-linked immunosorbent assay (ELISAs) for metalloproteinase derived type II collagen neoepitope, CIIM-Increased serum CIIM in subjects with severe radiographic osteoarthritis. Clin Biochem. 2011;44(5–6):423–9.